UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

February 4, 2000

DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter-S-RE-02

MEMORANDUM FOR John H. Thompson
Associate Director for Decennial Census

From: Howard Hogan Howard Hogan
Chief, Decennial Statistical Studies Division

Through: David Whitford and Magdalena Ramos
Decennial Statistical Studies Division

Subject: Research on Surrounding Block Rings for the Census 2000 Accuracy and Coverage Evaluation Processing

Prepared by: Glenn Wolfgang
Decennial Statistical Studies Division

1. Executive Summary

This research was motivated by a need to decide whether the Census 2000 Accuracy and Coverage Evaluation will extend the search for matches and duplicates to one ring or to two rings (as defined below) in order to correct geocoding error and thereby refine the precision of dual system estimates. This research investigates the potential impact on coverage measurement results from using one or two rings. That impact is one of many important considerations driving the decision. Post-production sample-wide alternative match results, along with surrounding block information, were used to identify additional possible matches and the ring in which they might be found. The possible matches were briefly evaluated for validity and completeness.

The number of new matches were cross-tabulated by ring (proximity to the sample cluster), by site, and by the density of new matches within the block. There were three main conclusions: (1) The bulk of new matches was in the first ring, but some could still be found outside even the second ring. Less than 0.5% of all P-sample records were in the second ring; about 7 to 8 times as many were in the first ring. (2) Rural areas, compared to urban sites, had fewer absolute new matches, but three times the gain in matches from extending to the second ring. Rural second-ring new possible matches were only about 0.1% of all P-sample records, but about 1.5% of all rural P-sample records. (3) Most of the new matches were found in a small number of blocks. Over half of the new matches were found in about one sixth of the blocks studied. The blocks that yielded the most new matches were sometimes in the second ring and even beyond.

The results were balanced with practical considerations regarding resources, timing, and development of operations, to yield a recommendation in favor of extending searches to the first ring but not the second ring.

2. Introduction and Purpose

Some of the Dress Rehearsal ICM cases classified as nonmatches, those never found among the census enumerations, were actually enumerated in the census but were simply hard to find because of errors in the location identities, called geocodes. Those geocoding errors theoretically would not affect bias in the final dual system estimates because they also should affect identification of duplication in census enumerations to a balancing degree. However, they were believed to increase the variance of the estimates.

It has been observed that, when geocoding error occurs, the correct geocode was often for a neighboring block, simply misplacing the designation of location from one block to an adjacent one. Searching among census enumerations in blocks adjacent to the sample cluster, the first surrounding ring, may net additional matches and corrected enumeration codes. Specifically, identifying among census records a match to a P-sample record not otherwise matched would refine the match data. Identifying E-sample cases that do not match P-sample cases but do match census cases in surrounding blocks refines the duplicate data. Together, those improvements would improve the precision of the estimates.

Extending the search to a second ring, blocks outside but adjacent to the first ring of surrounding blocks, could improve the precision even more. However, cost-effectiveness would diminish if search procedures were extended to the second ring. From some past experience, a large review workload and few resolved new matches were expected beyond the first ring. This study was developed and implemented quickly to help judge whether to search for geocoding error in either one or two rings. It was descriptive and exploratory rather than confirmatory.

One focus of the study was the potential for geocoding error from the second ring of blocks relative to the first. A secondary question was whether there are differences in certain sites (TEA values). A third focus of the research emerged with an observation pointing to a potential new strategy for focusing or targeting surrounding block searches more efficiently and effectively.

3. Data

This research uses several files as data sources:

3.1 New possible matches of ICM nonmatches to records in the latest CUF.

The file of new possible matches was generated by staff (David Word and Robert Nunziata) in the Population Division (POP). They began with a file of more than 15,000 Dress Rehearsal ICM persons who still had not been matched to census persons at the close of normal ICM production processing. This included search of the first ring of 17 block clusters identified for an extended search operation. They also used the latest version of the Census Unedited File (CUF). When records from the two files were found to have similar name and birth date, the data and identifiers, including geocodes, from both were combined onto a new possible match record.

The new match methodology was very different from that of production matching. The production matching was usually limited to the sample cluster. In only the seventeen clusters with the most severe signs of geocoding error, was the first ring of neighboring blocks also searched. The new matches were brought together from anywhere in the California or South Carolina sites, rather than just close to the sample block. Production matching entailed the most recent and sophisticated record linkage technology and a well-developed clerical review and coding procedure; much effort was put into accuracy. The simplified procedure used to find new matches was based on comparing only name and birth date, so the new matches included some false matches and missed some good matches. The new matches complemented production matches, but were still essentially exploratory, not confirmed, matches.

Possible new matches were classified into six useful quality types:

A Exact match on {first name, last name, month, day, and year (of birth)}.

B Exact match on first three characters of {first name, last name} & exact match on {month, day, and year (of birth)}.

C Exact match on {first name, last name} & on 2 of 3 date-of-birth elements.

D Exact match on first three characters of {first name, last name} & on 2 of 3 date-of-birth elements.

E Exact match on {first name, last name} & date-of-birth missing on CUF.

F Exact match on first three characters of {first name, last name} & date-of-birth missing on CUF.

3.2 GEO's list of surrounding blocks for sample clusters.

Geography division (GEO) developed an algorithm that identifies all the adjacent block codes for each block or block cluster. Each record in this file contained the identifiers of a sample cluster and one of its surrounding blocks.

3.3 GEO's list of surrounding blocks for target blocks.

Each record in this file held the identifiers for census blocks where possible matches reside (called target blocks) and their surrounding blocks. The list of target blocks was generated from the new possible match records in the first file above.

3.4 Extracts of CUF name/address and other data for 32 target blocks found to be in the second ring around the sample cluster and containing newly matched persons.

Extra variables were drawn from census records for a quick evaluation of the validity of new matches involving this small group of blocks.

3.5 Extracts of ICM name/address and other data for sample clusters corresponding to the 32 target blocks in the second ring.

Extra variables were drawn from survey records for a quick evaluation of the validity of new matches involving this small group of blocks.

4. Assumptions

This research was based on these underlying definitions, assumptions, and limitations:

4.1 The **first ring** of blocks surrounding a block or block cluster was understood to include all other blocks, excluding that cluster's blocks, which have boundaries that touch the block or block cluster at least at one point.

4.2 The **second ring** of blocks surrounding a block cluster was understood to include all blocks, excluding blocks in that cluster or its first ring, which have boundaries that touch a first ring block at least at one point.

4.3 A **target** block was the block of the census record newly matched to a P-sample case.

4.4 **Density** was the number of new matches found in a target block for one sample cluster's cases or an average over several cluster-target pairs.

4.5 Menominee site data were not in this study; new matches had not been produced there.

4.6 For the purpose of this research, water blocks were included in the first ring of surrounding blocks for both sample clusters and target blocks.

4.7 For the purpose of this research, surrounding blocks that crossover into other states or counties (separated by map file boundaries called TIGER partitions) were not essential to identifying substantial second ring contributions. GEO programs to find and list such blocks were not developed at the time the files were produced. As a result, it is possible that this research failed to identify a small number of target blocks in first or second rings.

4.8 Type F possible new matches were dropped from analysis because they were very likely false matches. See "6. Evaluation of New Matches" below for more information.

4.9 If only one new match was found in a given cluster-target pair, it was considered likely to be a false match or to not involve geocoding error.

4.10 If the new possible match had a partial household nonmatch code, it was considered not likely to involve geocoding error.

4.11 A rigorous, precise matching review was not conducted nor desired for the timing and scope of this research. Extensive field work and clerical review would have been essential to confirm or resolve uncertainty in publicly released data. The decision-oriented purpose prescribed a quick, minimal-resource comparison of the relative contribution of different types of target blocks to reducing geocoding error. From a strict scientific viewpoint these results should be replicated with finer controls and rigorous case review, if estimates were to be extrapolated from it or applied to other analyses.

4.12 Due to time constraints, this study did not pursue hypothesis testing nor try to assess the degree of impact of the second ring on variances.

4.12 This study did not explore census duplication due to geocoding error.

5. Procedure

5.1 The new match file was received from POP division [2/10/99].

5.2 A list of target blocks was extracted from the new match file and forwarded to GEO [with draft specifications 2/22/99].

5.3 GEO created a file listing surrounding blocks for target blocks [2/25/99].

5.4 GEO created a file listing surrounding blocks for sample clusters [3/3/99].

5.5 Possible new matches with the highest likelihood of being false matches or unrelated to geocoding error were dropped from the subsequent analysis. See 6.1 and 6.2 in "Evaluation of New Matches" below for detail.

5.6 A data set of block-level records was generated for sample clusters. It consisted of a record for each block in a sample cluster with retained new matches or surrounding blocks to those clusters. Each record contained cluster and block IDs as well as SRING, a variable identifying sample cluster (0) or adjacent block (1).

5.7 A data set of records was generated for the new match target blocks. It consisted of a record for each target block and each target block's surrounding block. Each record contained target/surrounding block ID and the paired cluster ID and TRING, a variable to distinguish a target block (3) from an adjacent block (2). When a block was the target for more than one cluster, it's target/surrounding block records were reproduced for each cluster.

5.8 When sample and target data set block identifiers matched (in a SAS merge of the two data sets above), the value of TIER=SRING-TRING+3 designated in which ring the target was found. TIER=0 meant the target block was in the cluster; a target block that did not connect with its cluster was considered "outside two rings."

5.9 To evaluate the quality of the new matches, name, address, and a few other data from evaluation (ICM) and census (CUF) were reviewed using records from the 31 clusters and 31 target blocks found in the second ring. See 6.3 in "Evaluation of New Matches" below for the summary of those findings.

5.10 Tallies of new matches cross-categorized by ring, site, and density levels were interpreted and used in calculating crude match rates. See "7. Results" for those findings.

6. Evaluation of New Matches

Ultimately the quality of this surrounding block research depended on whether the new match data were valid or true rather than spurious and on whether they were relevant to the purpose of correcting geocoding error.

6.1 POP provided the first means of evaluating possible new matches by classifying them into the quality types described in section 3.1. Type F possible new matches were very likely false matches. This was supported by a quick review of the new match file. While the first three letters of first and last names matched, full names were obviously different and birth date comparisons were useless. Some type F matches might have been true, but they would be a small percentage and hard to pick out. A total of 3060 type F new matches were dropped from the analysis; 4366 type A-E new matches remained.

6.2 When a target block had only one new match for cases from a given cluster, it was a sign that geocoding error was not involved. Geocoding error usually involved more than one address in a block and, moreover, whole households at those addresses. Movers, false matches, and persons with floating addresses might show up as single nonmatches. Only cursory investigation of such cases was done to verify that they were not geocoding error. However, 1682 new matches were dropped from the analysis because they were in target blocks with density equal to one; 2684 new matches in higher density blocks remained.

6.3 In eight new matches (before dropping low density targets), the target block was in the sample cluster. Since these particular cases were all in the dress rehearsal production review, they were referred again to NPC staff in order to learn more about the new match process. Not all materials used in the production review could be reassembled, but most of the cases were clearly false matches, and only a few were questionable. In the context of more than 54,000 production matches, eight or fewer false new matches were nearly insignificant, especially since all but three were dropped in the density test. This finding seemed to support the view that in blocks where many valid matches exist, there were relatively few false new matches that weren't weeded out by the process above.

6.4 At the other extreme of target remoteness, 29 new matches were found in 14 target blocks found in a different state than the sample cluster. Even if these were true matches, they were not likely to be due to geocoding error.

The high percentage of low-density targets beyond the second ring did suggest that the new match methodology provided a steady flow of false matches uniformly distributed over all records in all areas submitted to the matching. Geocoding error peaked in blocks close to the cluster and trailed off with distance.

This was consistent with trends suggested in following results (and worthy of finer testing or analysis): the farther a target was from the cluster, the more likely it was low density; and the lower the density of the target, the less likely it yielded corrections to geocoding error.

6.5 The greatest review of new matches was concentrated in the cluster-target pairs of the second ring. For each of 323 new matches initially identified in the second ring, various data from ICM and CUF sources were also reviewed. Three of those new matches were previously partial household nonmatches and should not have been counted as geocoding error; geocoding error was expected to be associated with whole households. Seven new matches involved an ICM record coded NU, for which a partial match probability was imputed and, to the extent that they were treated as matches, should have been counted as geocoding error. Three false matches were evident. Additional review of other census names in the target block and of other P-sample nonmatches in the sample cluster yielded 85 more possible matches missed by the new match methodology, except possibly a few that had been type F new matches. This missed-match effort was very superficial and should be interpreted as very rough approximation. But it did serve to show which way a rigorous extended search would go if conducted in these target blocks. For a variety of reasons, a few of the 323 cases should have been dropped from the new matches and many more should have been added. Such adjustments to this study's data were not made, however, because comparable numbers for new matches outside the second ring were not identified. Comparisons between the rings would not be valid with refinements only to the second ring. Indeed, most of the new matches were valid and perhaps up to 25% more matches could have resulted in production.

Addresses were also compared in this quick matching review. In 222 of the matches, addresses were similar enough to support the view that geocoding of the address was the source of error. In only one household were addresses clearly conflicting. The remaining were different styles of addresses where further investigation would be needed to confirm the addresses corresponded to the same location. One large block in a rural area had two matching addresses and 71 where differences in address style did not permit comparison. Still the consistent pattern of households matching from cluster to target strongly supported the validity of the new matches and that geocoding error was the reason for not finding the matches in production.

6.6 In summary, there were two good signs that this new match data picked out geocoding error: (1) all or most people in the ICM household matched those in the census household, and (2) the block had many (a high density of) new matches. The second ring displayed these traits often. The first ring new matches displayed the density trait also; moreover, it had a higher absolute frequency of new matches.

7. Results

The observed new matches should be understood in the context of the main production results. So Table 1 presents a summary of record counts for final 1998 Dress Rehearsal match codes along with, in the lower part, the further breakdown of new matches found for nonmatches after production was finished. The counts are also broken down by site: two large cities (Sacramento, CA and Columbia, SC), other mail-out/mail-back areas of SC, and rural areas of SC which were enumerated with an update/leave strategy.

Cells in the new match section of the table contain, after the counts of new matches, the number of target blocks in which those counts are found. Dividing the new match count by the number of targets, as the slash suggests, would give you the average density of new matches in the target block.

Table 1. Dress Rehearsal ICM Final Match Codes and Possible New Matches by Site

	Mail-Out/ Mail-back (TEA 1)			Update/ Leave (TEA 2)	
	Sac., CA	Col., SC	Other SC	Rural SC	TOTAL
Total Dress Rehearsal	36336	17810	12716	5394	72256
Deleted or Unresolved	1345	671	429	167	2612
Match in cluster	27094	14250	8765	3915	54024
Match in surrounding area	256	149	91	7	503
Nonmatch	7641	2740	3431	1305	15117
new matches / target	1002 / 128	515 / 88	915 / 158	255 / 40	2687 / 414
— in cluster	0	3 / 1	0	0	3 / 1
— in first ring	743 / 70	377 / 53	756 / 116	145 / 26	2021 / 265
— in second ring	139 / 9	35 / 5	64 / 12	85 / 6	323 / 32
— outside two rings	120 / 49	100 / 29	95 / 30	25 / 8	340 / 116

Two main results show up in Table 1 and a third on Table 2:

7.1 First ring versus second ring new matches: the bulk of new matches was in the first ring, but some good matches were still found in the second ring or beyond.

Comparing the rows in Table 1, new matches in the second ring (323) were less than half a percent of all P-sample records (72256, on the Total Dress Rehearsal line) and 2.1% of the dress rehearsal nonmatches (15117). First ring new matches (2021) were 2.8% of all P-sample records and 13.4% of the nonmatches.

For every second ring match, there were more than six first ring new matches, a {6.3:1} ratio. When including matches found in the surrounding block during the dress rehearsal production (503), the ratio of first to second ring matches was {7.8:1}; when adjusting for the missed new matches (using a factor of 1.25, based on the finding in 6.5) the ratio came to {7.5:1}.

Crude match rates were also computed by subtracting deleted or unresolved cases from the P-sample count (72256 - 2612 = 69644) in the base of the rate and by adjusting for missed matches and by adding in the first ring production matches as appropriate. The first ring search would thereby result in 3.63% more matches than without an extended search, and the second ring would contribute 0.58% more.

7.2 Sites: rural areas had fewer absolute new matches but a higher rate of gain in extending to the second ring.

Comparing the columns in Table 1, the contribution of the second ring differed by site. Sacramento and Columbia were quite similar, but the rest of SC TEA 1 had 10% more first-ring new matches. SC TEA 2 had a 23% fewer first-ring new matches and 20% more second-ring new matches than the TEA 1 sites. In particular, update/leave second-ring gains were 4.3 times the contribution relative to mail-back sites (1.5% versus 0.35% new matches out of all records). Crude match rates showed similar percentage increases.

Table 2. New Matches in Categories of Target Block Density

	Mail-Out/ Mail-back (TEA 1)			Update/ Leave (TEA 2)	
	Sac., CA	Col., SC	Other SC	Rural SC	TOTAL
2-9 new matches / target	364 / 104	241 / 73	468 / 133	119 / 35	1192 / 345
— in cluster	0	3 / 1	0	0	3/1
— in first ring	211 / 48	152 / 40	353 / 94	83 / 23	799 / 205
— in second ring	33 / 7	10 / 4	44 / 11	11 / 4	98 / 26
— outside two rings	120 / 49	76 / 28	71 / 28	25 / 8	292 / 113
10-19 new matches / target	172 / 12	116 / 10	199 / 16	39 / 3	526 / 41
— in cluster	0	0	0	0	0
— in first ring	172 / 12	116 / 10	175 / 14	29 / 2	492 / 38
— in second ring	0	0	0	10 / 1	10 / 1
— outside two rings	0	0	24 / 2	0	24 / 2
20+ new matches / target	466 / 12	158 / 5	248 / 9	97 / 2	969 / 28
— in cluster	0	0	0	0	0
— in first ring	360 / 10	109 / 3	228 / 8	33 / 1	730 / 22
— in second ring	106 / 2	25 / 1	20 / 1	64 / 1	215 / 5
— outside two rings	0	24 / 1	0	0	24 / 1

Table 2 presents the new match data broken down by one more variable, category of target block density. It illustrates a third main result regarding how density and distance from the sample cluster and site interacted.

7.3 Density: Most of the new matches were found in a small number of blocks.

Comparing the sections of Table 2, there are in any ring a few blocks that had the bulk of the new matches, the index of geocoding errors. The 6.8% of blocks that had the highest density (20+) of new matches (out of the total from Table 1), had 36% of all new matches. Or, the 16.7% high density (10+) target blocks had 55.7% of the new matches. Six specific target blocks (1.5%) in the second ring (among 10+) would provide 8.4% of the new matches. High density blocks yielded the biggest payoff for processing geocoding error.

8. Conclusions

This study was motivated by a practical operation question: Should we extend our search for geocoding error to one ring or two rings around the sample cluster? Even in the context of empirical data, the practical concerns associated with the question must be kept in perspective.

Indeed, while the technical benefit of reducing variance is very important, the operational difficulties weighed heavily against the second ring. GEO might not be able to program identification of the second ring in time or might have difficulties displaying them on maps. Additional staffing to handle the increased workload would be an extraordinary strain for FLD and DPD. Changes or additions to the imaging process might be a problem for DSCMO. Changes to the existing procedures would be a burden and risk for everyone. Other technical and operational difficulties might arise if not handled correctly.

The results boil down to a few points: most new matches were in the first ring; rural areas may have had fewer absolute new matches but had a higher rate of gain in extending to the second ring; the greatest gain in new matches was in a few blocks where there was a concentration of matched households, apparently geocoding error. That was true regardless of the ring.

Balancing the operational burdens against the technical gains, we may simply have to settle for what can be managed, namely one ring, in Census 2000. We take comfort in the first ring giving the majority of gain in cleaning up geocoding error. We look forward to designing a more incisive extended search procedure for the census beyond 2000 and believe the clues are in this study to suggest a direction in doing so.

The last observation about target density suggests that there should be a targeting procedure, designed to be both timely and unbiased, that makes searches (even beyond the first ring) more efficient and effective. The issue of extending the search beyond the first ring may be viewed as independent of whether there is any targeting of blocks within those rings (second stage of selection after the targeting of clusters for TES). On the other hand, we might target TES clusters without reference to surrounding rings. Research to develop balanced and efficient targeting of geocoding error, perhaps using new possible match data like that used in this study, would be an advance for Census 2010.

cc: DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
A.C.E. Implementation Team
Statistical Design Team Leaders